Math 158 – Linear Models

Spring 2018

Jo Hardin

iClicker Questions

Review

1. The p-value is defined as:

(a) The probability that H0 is true.

(b) The probability that H0 is true given the observed data.

(c) The probability of observing the data.

(d) The probability of observing the data given H0 is true.

(e) I’ve never encountered the definition of the p-value.

Review

2. In order to draw a causative conclusion, we typically need:

(a) a randomization experimental design

(b) a random sample

(c) a very large sample

(d) a good theory for why the causation should exist

(e) I have no idea

Review

3. Power is:

(a) The probability that H0 is true.

(b) The probability that H0 is false.

(c) The probability of observing our data.

(d) The probability of rejecting H0.

(e) I know power is good, but I can’t remember why.

SLR

4. True or False:

The equation

y = 5x + 2

represents the equation of a line with slope 2.

(a) True.

(b) False.

SLR

5. True or False:

The lines

y = 3x + 6 and

y = -1/3x + 2

are perpendicular.

(a) True

(b) False

SLR

6. Does the equation

12x - 8y + 4 = 0

determine a linear function of x?

(a) It does determine a linear function.

(b) It does not determine a linear function.

SLR

7a. Consider switching the roles of X and Y

model 1: Y ~ X

model 2: X ~ Y

(a) b1(model1) = b1(model2)

(b) b1(model1) = 1 / b1(model2)

(c) b1(model1) = - 1 / b1(model2)

(d) b1(model1) = - b1(model2)

(e) b1(model1) ≠ b1(model2)

SLR

7b. If the null hypothesis (that β1=0) is true (and b1=0), our SSE reduces to:

(a) ∑ (i - ) 2

(b) ∑ (Yi – i) 2

(c) ∑ (Yi - ) 2

SLR

8. The regression technical conditions include:

(a) The Y variable is normally distributed

(b) The X variable is normally distributed

(c) The residuals are normally distributed

(d) The slope coefficient is normally distributed

(e) The intercept coefficient is normally distributed

SLR

9. What happens if the technical assumptions are violated?

(a) The inference goes askew.

(b) The line doesn’t fit.

(c) The sample isn’t representative.

(d) It isn’t appropriate to draw a causation conclusion.

SLR

10. Other than SLR, how can we evaluate the relationship between

belonging to a Greek organization and GPA?

(a) chi-squared test of independence

(b) t-test for comparing means of independent groups

(c) t-test for comparing means of dependent groups

(d) z-test for comparing means of independent groups

(e) chi-squared test of goodness-of-fit

SLR

11. To test if there is convincing evidence that the slope of the regression line between debt and income is different from zero,

what are the appropriate hypotheses?

(a) H0: b0 = 0

Ha: b0 ≠ 0

(b) H0: b1 = 0

Ha: b1 ≠ 0

(c) H0: β0 = 0

Ha: β0 ≠ 0

(d) H0: β1 = 0

Ha: β1 ≠ 0

SLR

12. Which of the below is the correct?

(to assess a negative relationship between absences and GPA)

Coefficients:

Estimate Std. Err t-val Pr(>|t|)

(Inter) 3.47 0.42 8.35 7.8e-09

Absences -0.016 0.01 -1.9 0.068

(a) H0:b1=0 Ha:b1≠0 p-value=0.068

(b) H0:β1=0 Ha:β1 < 0

p-value = 0.068/2 = 0.034

(c) H0:β1=0 Ha:β1 ≠ 0

p-value = 0.068/2 = 0.034

(d) H0:b1=0 Ha:b1 < 0

p-value = 0.068/2 = 0.034

(e) H0:β1=0 Ha:β1 ≠ 0 p-value=7.8x10-9

SLR

13. We created a 95% confidence interval for the **mean GPA** given 10 absences to be (3.20, 3.42). What is the correct interpretation?

(a) There is a 95% chance that the mean GPA of students with 10 absences is between 3.20 and 3.42.

(b) 95% of GPA averages (for students with 10 absences) are between 3.20 and 3.42.

(c) 95% of GPAs (for students with 10 absences) are between 3.20 and 3.42.

(d) We are 95% confident that the true mean GPA (for students with 10 absences) is between 3.20 and 3.42.

(e) 95% of our intervals will have a mean GPA between 3.20 and 3.42.

SLR

14. We created a 95% prediction interval for an **individual GPA** given 10 absences to be (3, 3.62). What is the correct interpretation?

(a) There is a 95% chance that the mean GPA of students with 10 absences is between 3 and 3.62.

(b) 95% of GPA averages (for students with 10 absences) are between 3 and 3.62.

(c) 95% of GPAs (for students with 10 absences) are between 3 and 3.62.

(d) We are 95% confident that the true mean GPA (for students with 10 absences) is between 3 and 3.62.

(e) 95% of our intervals will have a mean GPA between 3 and 3.62.

SLR

15. A confidence interval for a mean response at Xh:

(a) has a constant (width) for any value of Xh.

(b) increases in width as Xh increases.

(c) decreases in width as Xh increases.

(d) increases in width as Xh gets farther from .

(e) decreases in width as Xh gets farther from .

SLR

16. A prediction interval for an individual observation at Xh:

(a) has a constant (width) for any value of Xh.

(b) increases in width as Xh increases.

(c) decreases in width as Xh increases.

(d) increases in width as Xh gets farther from .

(e) decreases in width as Xh gets farther from .

SLR

17. R2 for the regression line for predicting GPA based on absences is 91.31%.

91.31% of

(a) GPAs can be accurately predicted by absence.

(b) variability in predictions of GPA is explained by absence.

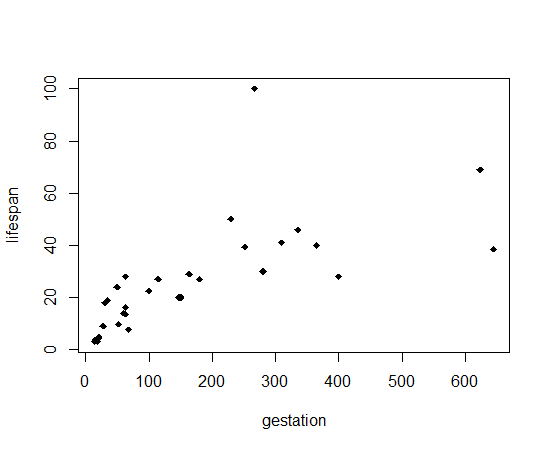
(c) variability in predictions of absences is explained by GPA.

(d) variability in GPA is explained by absences.

(e) variability in absences is explained by GPA.

SLR

18. Which SLR assumption is violated?



(a) linearity

(b) constant variance of errors

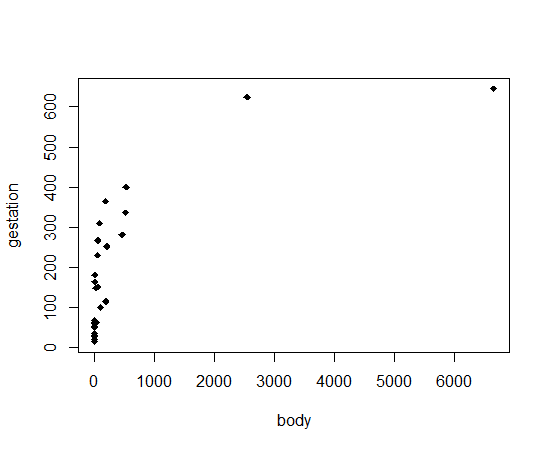
(c) independent errors

(d) normal errors

(e) outliers

SLR

19. Which SLR assumption is violated?



(a) linearity

(b) constant errors

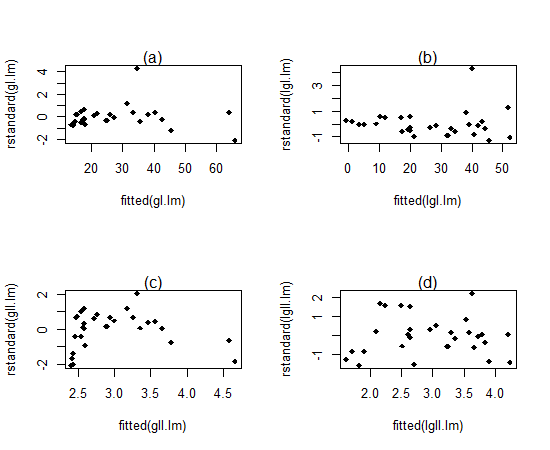
(c) independent errors

(d) normal errors

(e) outliers

SLR

20. Which residuals are “best” ?



SLR

21. Why do we plot the residuals versus **fitted** instead of **explanatory**?

(a) because the residuals are correlated with the fitted variable.

(b) because the residuals are correlated with the explanatory variable

(c) because we want to be able to extend the model to incorporate more variables

(d) because the explanatory variable is a linear combination of the fitted values

SLR

22. Why do we plot the residuals versus **fitted** instead of **observed**?

(a) because the residuals are correlated with the fitted variable.

(b) because the residuals are correlated with the observed response

(c) because we want to be able to extend the model to incorporate more variables

(d) because the observed variable is a linear combination of the fitted values

SLR

23. Consider the following regression model:

If we go from X=20 to X = 40 in the model, we can interpret as

(a) E[Y] is larger by an amount of 20

(b) median[Y] is larger by a factor of .

(c) E[Y] is larger by an amount of .

(d) median[Y] is larger by a factor of.

SLR

24. Consider the following regression model:

If we go from X=20 to X = 40 in the model, we can interpret as

(a) E[Y] is larger by an amount of 20

(b) median[Y] is larger by a factor of .

(c) E[Y] is larger by an amount of .

(d) median[Y] is larger by a factor of.

SLR

25. Consider the following regression model:

If we go from X=20 to X = 40 in the model, we can interpret as

(a) E[Y] is larger by an amount of 20

(b) median[Y] is larger by a factor of .

(c) E[Y] is larger by an amount of .

(d) median[Y] is larger by a factor of.

SLR

26. Consider the following regression model:

If we go from X=20 to X = 40 in the model, we can interpret as

(a) E[Y] is larger by an amount of 20

(b) median[Y] is larger by a factor of .

(c) E[Y] is larger by an amount of .

(d) median[Y] is larger by a factor of.

SLR

27. The intercept in the multiple regression model

a. should be excluded if one explanatory variable has negative values.

b. determines the height of the regression line.

c. should be excluded because the population regression function does not go through the origin.

d.is statistically significant if it is larger than 1.96.

SLR

28. The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill ($) and home size (sq ft).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regression equation:**  Annual bill = 0.55 \* Home size + 15 | | | | |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| Constant | 15 | 3 | 5.0 | 0.00 |
| Home size | 0.55 | 0.24 | 2.29 | 0.024 |

What is the 99% confidence interval for the slope of the regression line?

(A) 0.25 to 0.85   
(B) 0.02 to 1.08   
(C) -0.08 to 1.18   
(D) 0.20 to 1.30   
(E) 0.30 to 1.40

SLR

29. The confidence interval for a single coefficient in a multiple regression

a. makes little sense because the population parameter is unknown.

b. should not be computed because there are other coefficients present in the model.

c. contains information from other hypothesis tests (here only ).

d. should only be calculated if the regression R2 is large.

Linear Algebra

30. Calculate

(a)

(b)

(c)

(d) None of the above

(e) This matrix multiplication is impossible

Linear Algebra

31. If A and B are both 2x3 matrices, which of the following is not defined?

(a) A+B

(b) AtB

(c) BA

(d) ABt

(e) More than one of the above

Linear Algebra

32. If A = and

B = , what is the (3,2)-entry of AB?

(a) 0

(b) 1

(c) 3

(d) 4

(e) 8

Multiple Linear Regression

33. Which of the below correctly describes the roles of variables in this regression model?

Est SE t value Pr(>|t|)

(Intercept) 197.96 59.20 3.34 0.0058

volume 0.71 0.06 11.67 0.0000

coverpb -184.05 40.49 -4.55 0.0007

(a) response: weight

explanatory: volume, paperback cover

(b) response: weight

explanatory: volume, hardcover cover

(c) response: volume

explanatory: weight, cover type

(d) response: weight

explanatory: volume, cover type

MLR

34. An econometrician is interested in evaluating the relation of demand for building materials to mortgage rates in LA and SF.

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

holding constant the effect of city, each additional increase of 1% in the mortgage rate would lead to an estimated average \_\_\_\_\_\_\_\_ in the mean demand.

(a) predicted $500 more per capita

(b) predicted $500 less per capita

(c) predicted $5 more per capita

(d) predicted $5 less per capita

MLR

35. Referring to

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

the effect of living in LA rather than SF is a \_\_\_\_\_\_\_\_ demand by an estimated \_\_\_\_\_\_\_\_ holding the effect of mortgage rate constant.

(a) larger; $800 per capita

(b) smaller; $800 per capita

(c) larger, $8 per capita

(d) smaller, $8 per capita

MLR

36. Referring to

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

the fitted model for predicting demand in LA is \_\_\_\_\_\_\_\_.

(a) 10 + 5X1

(b) 10 + 13X1

(c) 15 + 8X2

(d) 18 + 5X1

MLR

37. Referring to

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

the fitted model for predicting demand in SF is \_\_\_\_\_\_\_\_.

(a) 10 + 5X1

(b) 10 + 13X1

(c) 15 + 8X2

(d) 18 + 5X1

MLR

38. A dummy variable (0,1) is used as an explanatory variable in a regression model when:

(a) the variable involved is numerical.

(b) the variable involved is categorical.

(c) a quadratic relationship is suspected.

(d) when two explanatory variables interact.

MLR

39. If a categorical explanatory variable contains three categories, then \_\_\_\_\_\_\_\_\_ dummy variable(s) will be needed to uniquely represent these categories.

(a) 1

(b) 2

(c) 3

(d) 4

MLR

40. An interaction term in a multiple regression model may be used when:

(a) the coefficient of determination is small.

(b) there is a quadratic relationship between the response and explanatory variables.

(c) neither one of two explanatory variables contribute significantly to the regression model.

(d) the relationship between X1 and Y changes for differing values of X2.

MLR

41. Referring to

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

to test whether there is a location effect on demand, one would use:

(a) an F test on the significance of the whole regression model.

(b) a t test on the significance of β1.

(c) a t test on the significance of β2.

(d) None of the above.

MLR

42. The F statistic for testing the entire regression model can be expressed as:

(a) SSR/SSE.

(b) MSE/MSR.

(c) MSR/MSE.

(d) MSR/SST.

MLR

43. The adjusted R2 is "adjusted for" the:

(a) number of predictors only.

(b) sample size only.

(c) number of predictors and the sample size.

(d) None of the above.

MLR

44. In a multiple regression model, which of the following is correct regarding the value of the adjusted R2?

(a) It can be negative.

(b) It must be positive.

(c) It must be larger than the coefficient of multiple determination (R2).

(d) It can be larger than 1.

MLR

45. Which of the following is NOT an assumption for the multiple regression model?

(a) Positive autocorrelation of error terms.

(b) Normality of error terms.

(c) Independence of error terms.

(d) Constant variation of error terms.

(e) At any combination of values of the explanatory variables, the error terms have a mean of 0.

MLR

46. It is often a good idea to compute R2adj, the adjusted R2, rather than just R2. This is done in order to

(a) avoid overestimating the importance of the explanatory variables.

(b) correct any problems that may arise from any of the model assumptions being violated.

(c) avoid having a multiple regression model that contains too many variables.

(d) avoid conducting t tests for each explanatory variable.

MLR

47. An application of the multiple regression model generated the following results involving the F test of the overall regression model:

p-value = 0.0012, R2 = 0.67, s = 0.076.

The null hypothesis, which states that none of the explanatory variables are significantly related to the response variable, should be rejected, at the 0.05 level of significance.

A) True

B) False

MLR

48.

Which test should be used to test the significance of the higher order terms (and )?

H0:

Ha: At least one of and does not equal 0.

A) the overall F test.

B) the R2

C) the nested F test.

D) the t test.

MLR

49. In a nested F test, the reduced model corresponds to the null hypothesis being true.

(a) True

(b) False

MLR

50. In a nested F test, the full model corresponds to the alternative hypothesis being true.

(a) True

(b) False

MLR

51. Each of the following linear hypothesis can be tested using the nested F-test with the exception of

(a) β2 = 1 and β3 = β4 / β5

(b) β2 = 0

(c) β1 + β2 = 1 and β3 = -2β4

(d) β0 = β1 and β1 = 0

MLR

52. Regressing Y on 100 explanatory variables X1, X2, …, X100, R2 = 0.87.

What is the most appropriate way to interpret R2?

(a) 87 of the explanatory variables in the model are capable of accurately predicting Y.

(b) We will accurately predict Y 87% of the time.

(c) 87 of the explanatory variables are statistically significant, while the remaining 13 variables should be removed from the model.

(d) 87% of the variation in the response variable can be explained by the explanatory variables.

MLR

53. Based on data from 9 days, a multiple regression was fit to the Dow Jones Index on the following 7 explanatory variables:

i. high temp

ii. low temp

iii. 1 if sunny

iv. 1 if Yankees won

v. # runs Yankees scored

vi. 1 if Mets won

vii. # runs Mets scored

What do you think of R2?

(a) It will be low because none of the variables are likely to predict the Dow.

(b) It will be high because that combo of variables is likely to predict the Dow.

(c) It will be low for other (mathematical) reasons.

(d) It will be high for other (mathematical) reasons.

MLR

54. In case of multiple regression the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is the proportion of variation in the response variable that is explained by the combination of explanatory variables.

(a) coefficient of correlation

(b) coefficient of partial determination

(c) coefficient of regression

(d) coefficient of multiple determination

MLR

55. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_measures the proportion of variation in the response variable that is explained by each explanatory variable holding all other explanatory variables constant.

(a) coefficient of correlation

(b) coefficient of partial determination

(c) coefficient of regression

(d) coefficient of multiple determination

COINS:

$ ~ num coins, slope coef will be:

(a) positive

(b) negative

(c) zero

(d) can’t tell

COINS:

$ ~ low coins, slope coef will be:

(a) positive

(b) negative

(c) zero

(d) can’t tell

COINS:

$ ~ num coins + low coins,

slope coef on low coins will be:

(a) positive

(b) negative

(c) zero

(d) can’t tell

56. For a multiple regression model, the computer output shows that the simple correlation coefficient between the response variable and one of the explanatory variables is 0.99. This result indicates that most likely the problem of multicollinearity exists in this model.

(a) TRUE

(b) FALSE

MLR

57. Suppose that in a multiple regression the F is significant, but none of the t-ratios are significant. This means that:

(a) multicollinearity may be present

(b) residuals are not independent

(c) the regression is good

(d) a nonlinear model would be a better fit

(e) none of the above

MLR

58. All of the following are possible effects of multicollinearity EXCEPT:

1. SE of coefficients may be larger than expected
2. the signs of the coefficients may be opposite of what is expected
3. a significant F ratio with t ratios not significant
4. removal of one data point may cause large changes in the coefficient estimates
5. the VIF is zero

MLR

59. For a quiz on 100 topics

(you know nothing):

Kelly knows 85 topics.

Jamie knows 75 topics.

Parker knows 55 topics.

Riley knows 45 topics.

Who should you choose to help you answer the questions?

(a) Kelly

(b) Jamie

(c) Parker

(d) Riley

(e) can’t tell

MLR

60. Who do you want to choose next?

Kelly knows 85 topics.

Jamie knows 75 topics.

Parker knows 55 topics.

Riley knows 45 topics.

(a) Jamie

(b) Parker

(c) Riley

(d) depends on overlap with Kelly

(e) depends on overlap with Jamie, Parker, and Riley

MLR

61. If you can pick two people, who do you pick?

Kelly knows 85 topics.

Jamie knows 75 topics.

Parker knows 55 topics.

Riley knows 45 topics.

(a) Kelly plus person who overlaps least

(b) Kelly plus person who overlaps the most

(c) The two with the least overlap

(d) The two with the most overlap

(e) The two who have the largest union.

MLR

62. Which of following is true?

(a) Influential points always reduce R2.

(b) High leverage points always reduce R2.

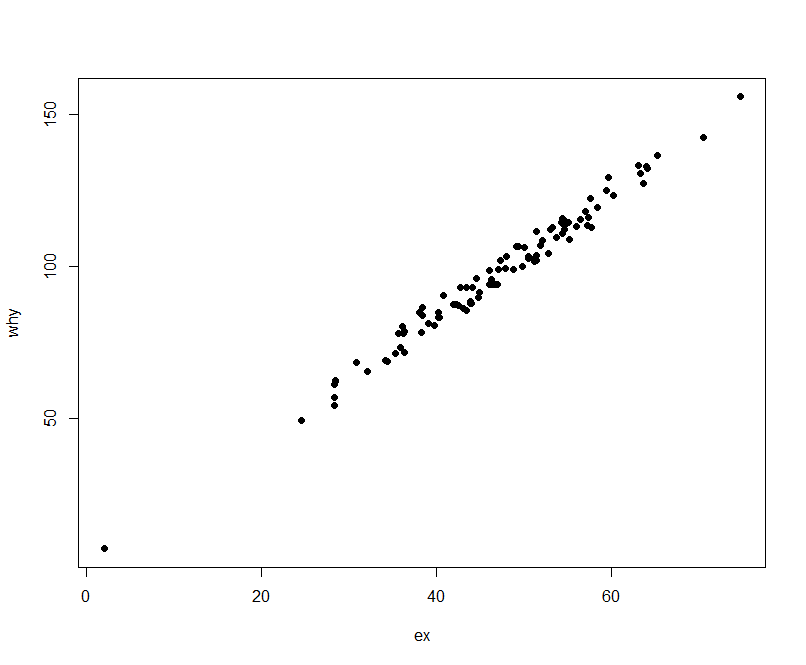
(c) All outliers are influential points.

(d) When the data set includes an influential point, the relationship between x and y is nonlinear.

(e) None of the above.

MLR

63. Which of the below best describes the outlier?



(a) influential

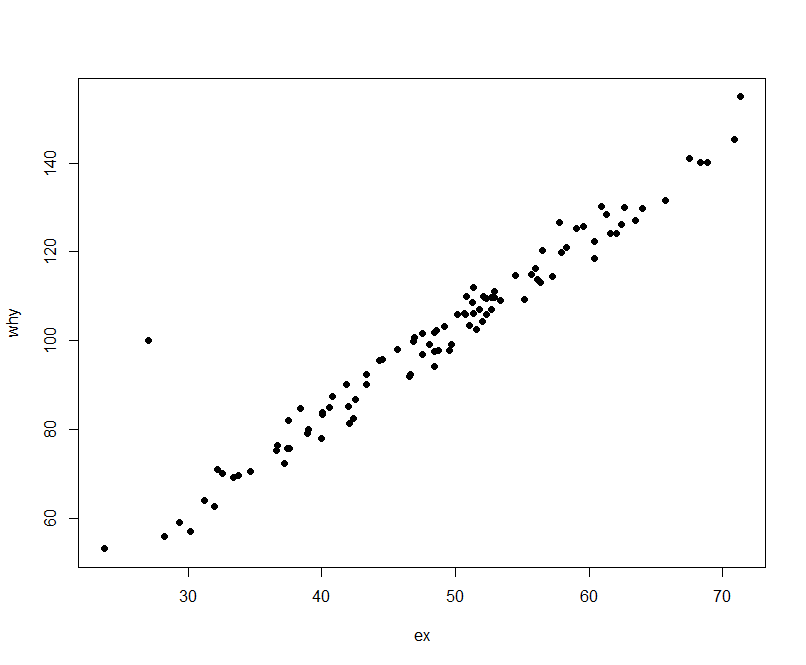
(b) low leverage

(c) high leverage

(d) none of the above

MLR

64. Which of the below best describes the outlier?



(a) influential

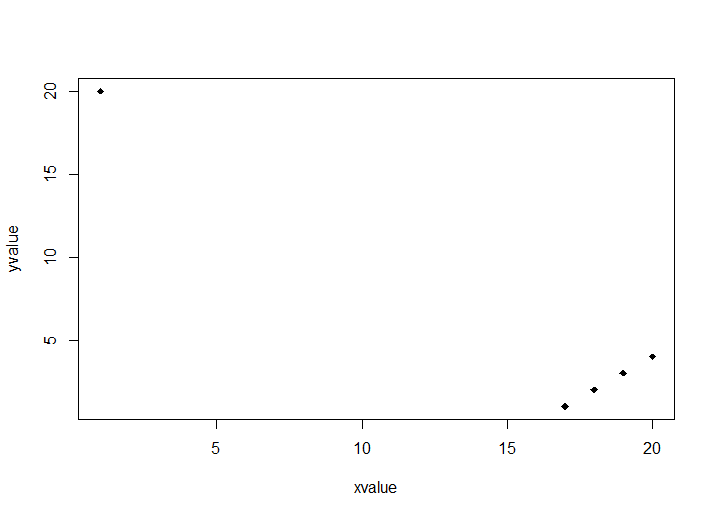
(b) low leverage

(c) high leverage

(d) none of the above

MLR

65. 1 obs in top left, 25 each in bottom right. r (correlation) is:



(a) 0.9-0.99

(b) 0.7-0.89

(c) 0.4-0.69

(d) 0 – 0.39

(e) < 0

MLR

65. Why does a case with large leverage have only the *potential* to be influential?

(a) If the sample size is large, all outliers will be mitigated.

(b) If the response at that point is consistent with the model given by the other points, the model won’t change.

(c) It may influence the coefficients but not the residuals.

(d) It doesn’t, high leverage points will always be influential.

MLR

66. To check whether the i have homogeneous variance:

(a) leverage plot

(b) Cook’s Distance plot

(c) DFBETAS plot

(d) Residual plot

(e) VIF plot

MLR

67. To check whether the regression is being unduly influenced by the 11th observation.

(a) leverage plot

(b) Cook’s Distance plot

(c) DFBETAS plot

(d) Residual plot

(e) VIF plot

MLR

68. To check whether the regression on X3 is really linear (as the model states).

(a) leverage plot

(b) Cook’s Distance plot

(c) DFBETAS plot

(d) Residual plot

(e) VIF plot

MLR

69. Is there an observation that does not seem to fit the model?

(a) leverage plot

(b) Cook’s Distance plot

(c) DFBETAS plot

(d) Residual plot

(e) VIF plot

Shrinkage Methods

70. Recall that n is the size of the data set and p is the dimension of the coefficient vector.

What is the size of the matrix that gets inverted in ridge regression?

(a) p x p

(b) n x n

(c) np x np

(d) n2 x n2

Shrinkage Methods

71. As λ 🡢 0 (in ridge regression or lasso)

(a) RR coefficients 🡢 ∞

(b) RR coefficients 🡢 0

(c) RR coefficients 🡢 LS coefficients

Shrinkage Methods

72. As λ 🡢 ∞ (in ridge regression or lasso)

(a) RR coefficients 🡢 ∞

(b) RR coefficients 🡢 0

(c) RR coefficients 🡢 LS coefficients

Shrinkage Methods

73. The main motivation of ridge regression is that it:

(a) minimizes bias

(b) minimizes variance

(c) minimizes (bias and variance)

(d) maximizes bias

(e) maximizes variance

Shrinkage Methods

74. The best procedure to assess multicollinearity is

(a) to examine the correlation matrix. (b) to examine the pairs plot

(c) to use the variance inflation factor. (d) to use ridge regression.

Shrinkage Methods

75. The lasso outperforms ridge regression at the cost of an increase in model complexity.

(a) Always

(b) Sometimes

(c) Never

Shrinkage Methods

76. Which of the following is associated with the procedure to get regression parameter estimates with smaller standard errors

(a) piece-wise regression

(b) polynomial regression

(c) ridge regression

(d) stepwise regression

(e) lasso regression

Smoothing

77. With step basis function, C0 through CK,

(a) all K+1 functions can be used as explanatory variables.

(b) any K functions can be used as explanatory variables

(c) C1 through CK should be used as explanatory variables

(d) any K-1 functions can be used as explanatory variables

(e) C2 through CK should be used as explanatory variables

Smoothing

78. In the step function model, how is interpreted? (On the basis function I(cj <= X < cj+1) .)

(a) the value of Y over cj <= X < cj+1

(b) the average value of Y over

cj <= X < cj+1

(c) the increase in Y from X < c1 to

cj <= X < cj+1

(d) the increase in average Y from

X < c1 to cj <= X < cj+1

Smoothing

79. In a polynomial model, what is the jth basis function?

(a) X

(b) X2

(c) Xj-1

(d) Xj

(e) Xj+1

Smoothing

80. What mathematics gives the needed information in order to perform inference for any basis function model?

(a) calculus

(b) Lagrange multipliers

(c) linear algebra

(d) some of the above

(e) none of the above

Regression Splines

81. Fitting separate polynomial models locally can be problematic because

(a) the variability at the extremes is high

(b) the higher order derivatives may not be continuous

(c) the step function may not be continuous

(d) the model can be numerically unstable because the explanatory variables are highly correlated

Regression Splines

82. T o fit a piecewise linear function with 3 knots, continuous at the nodes, one needs

(A) 3 bases functions

(B) 4 bases functions

(C) 5 bases functions

(D) 6 bases functions

Regression Splines

83. How many degrees of freedom do we have left when fitting a cubic regression spline with K knots?

(a) K

(b) K + 3

(c) K + 4

(d) n – K

(e) n – K – 4

Local Regression

84. Why is it difficult to extend local regression (loess) to higher dimension?

(a) in high dimensions all points are far from one another

(b) inverting matrices with p > 2 is computationally difficult

(c) distance is difficult to compute in high dimensions

Local Regression

85. Tricubic weight function:

(a) give decreasing (as a function of d(xi, x0) ) and non-zero weights for ALL values in training set

(b) cannot be computed for s > 1 “proportion” of the data

(c) can only be computed using Euclidean distance

(d) can sometimes be negative

(e) none of the above